

MULTILEVEL-ANALYSE

18 juni 2024

Training O + S

Elmar Jansen (elmar@elmarjansen.nl)

VANDAAG

1. Terugblik
2. Geneste Data
3. Fixed Effects Model (met dummies)
4. Multilevel: Random Intercept Model
5. Multilevel: Random Slope Model

DE KOMENDE WEKEN

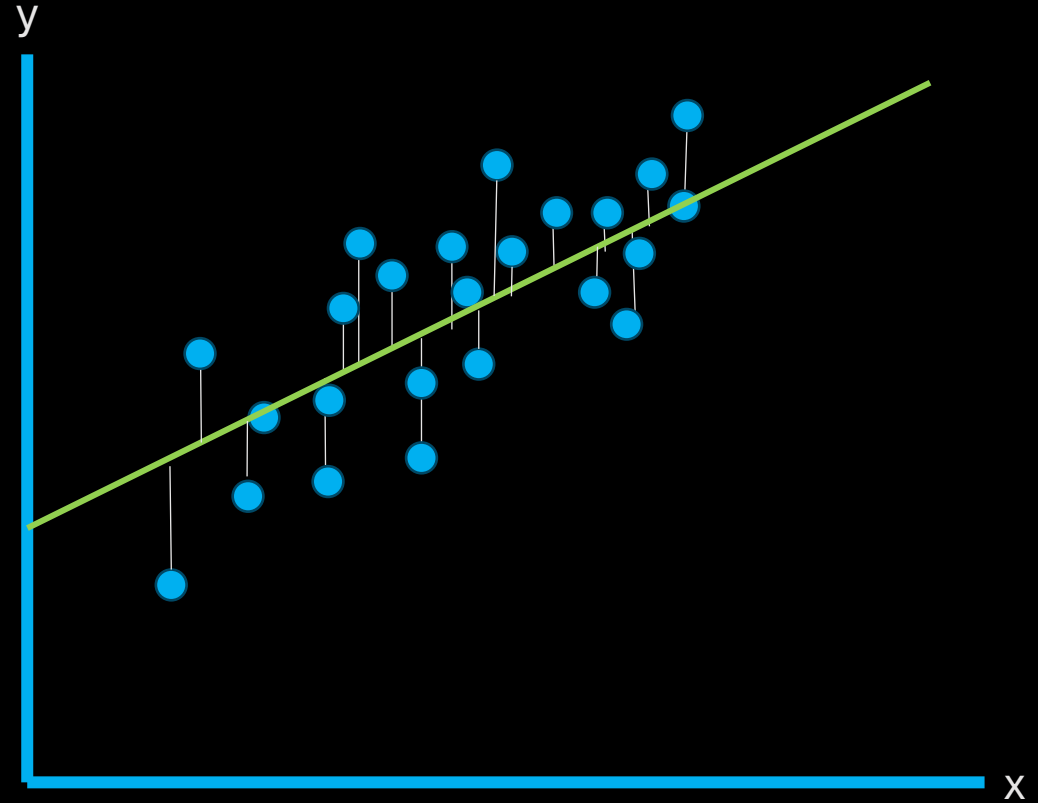
Bijeenkomst	Onderwerp
Dinsdag 14 mei	Lineaire regressie: de basis
Dinsdag 21 mei	Lineaire regressie vervolg: assumpties en controleren
Donderdag 30 mei	Interacties en dummy-variabelen
Dinsdag 4 juni	Logistische Regressie
Dinsdag 11 juni	Multilevel-analyse



TERUGBLIK

LINEAIRE REGRESSIE

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \varepsilon_i$$

DANGER

8 GEVAREN VAN REGRESSIE

DANGER

DANGER!!

1



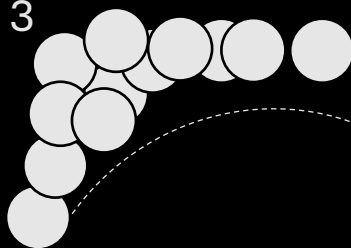
Schijnverband

2



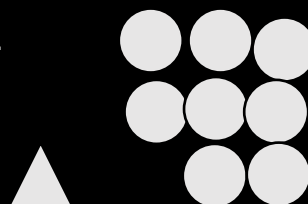
Wederkerigheid /
Simultaniteit

3



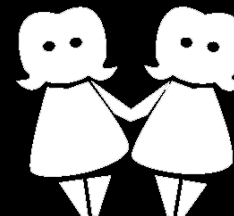
Non-Lineairiteit

4



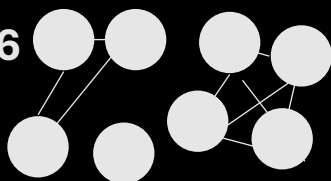
Extreme waarden

5



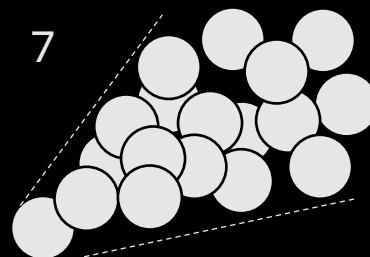
Multicollineariteit

6



Niet onafhankelijke
residuen

7



Heteroskedasticiteit

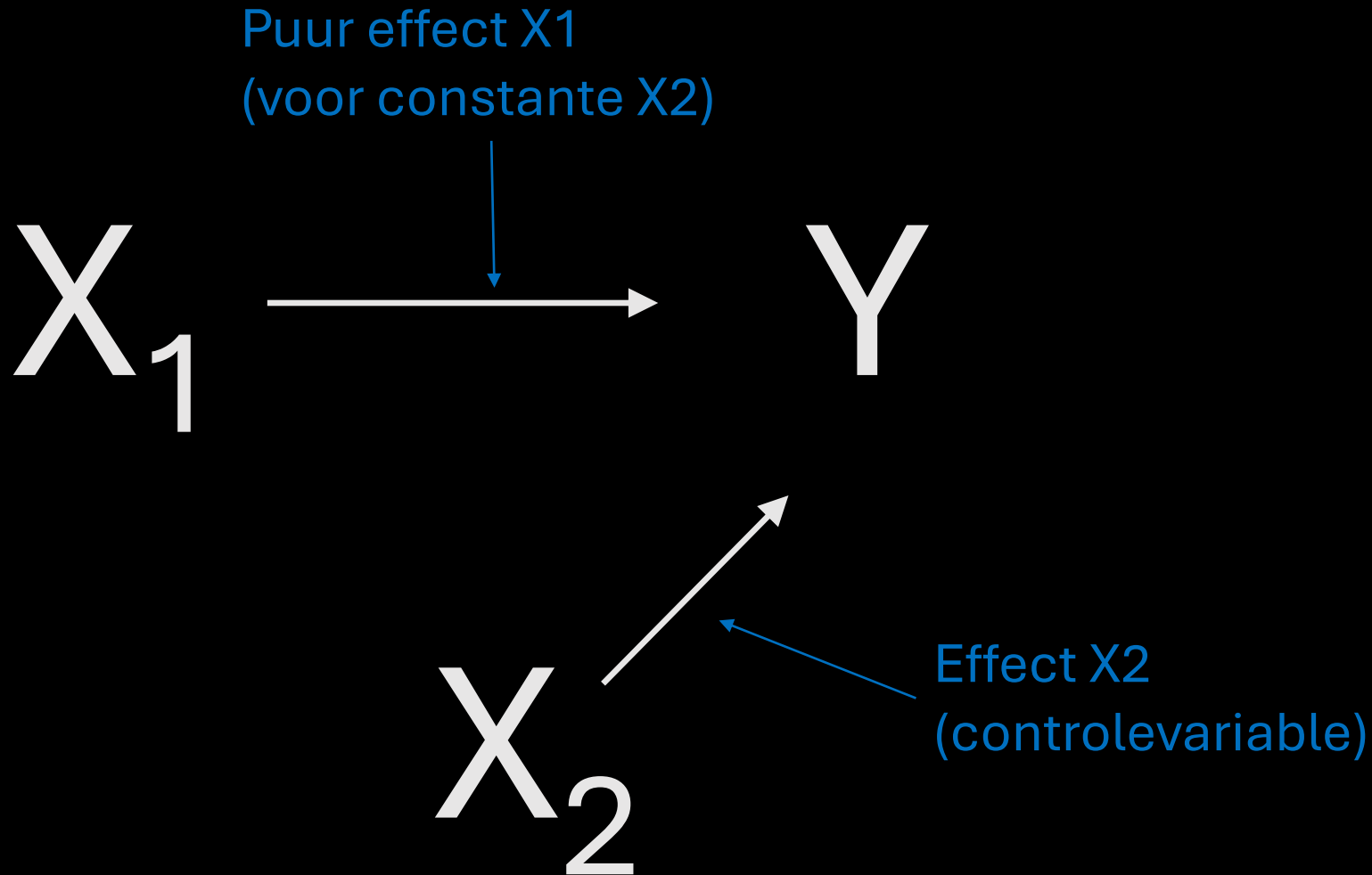
8



Non-normaliteit
van residuen

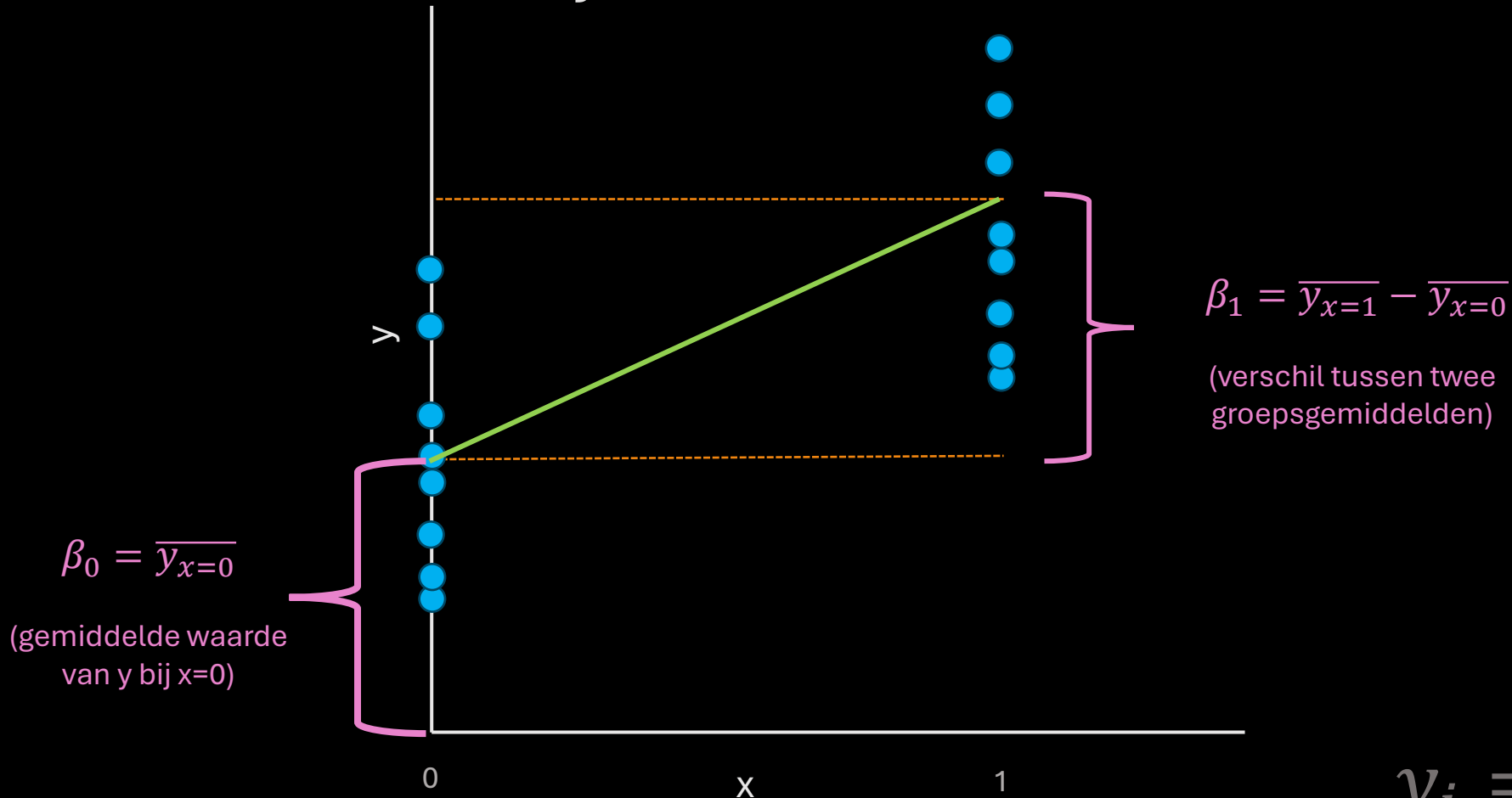
CONTROLLEREN

Door onafhankelijke variabele X_2 toe te voegen aan het model krijgen we het effect van X_1 **constant houdend voor X_2** (en viceversa)



DUMMY-VARIABELE

Dichotome variabele met waarden 0 en 1
als onafhankelijke variabele



$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

CATEGORIALE ONAFHANKELIJKE VARIABELEN

Categorie A

$$CatB_i = 0$$

$$CatC_i = 0$$

Categorie B

$$CatB_i = 1$$

$$CatC_i = 0$$

Categorie C

$$CatB_i = 0$$

$$CatC_i = 1$$



Baseline-categorie

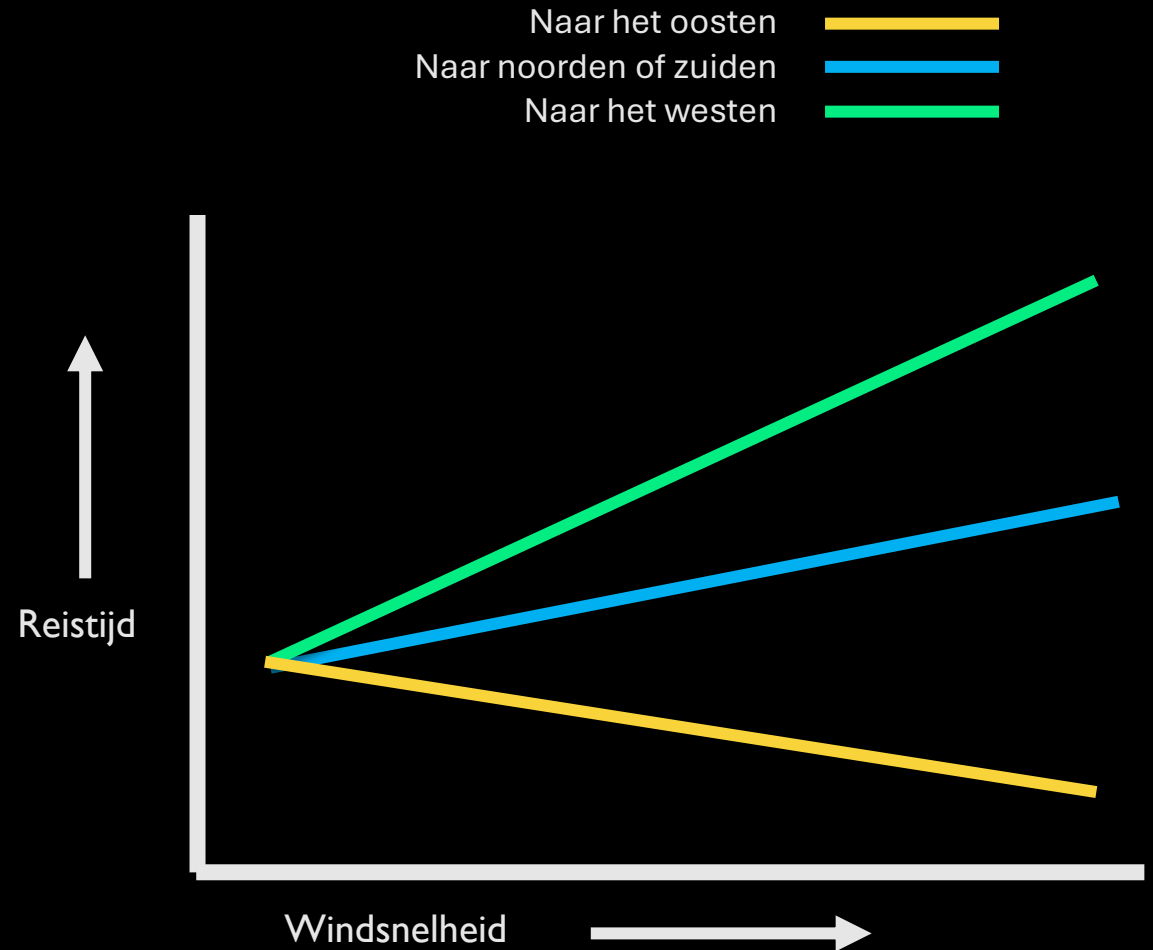
INTERACTIE-EFFECT

Een effect

van een variabele

op het effect van

een andere variabele



INTERACTIE-EFFECT IN REGRESSIE

Voeg ook altijd het
“main”-effect van
beide variabelen toe!

De interactie is de
vermenigvuldiging
tussen beide variabelen

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$$

Als je herschikt, zie je dat het effect van X1 nu afhankelijk is van X2:

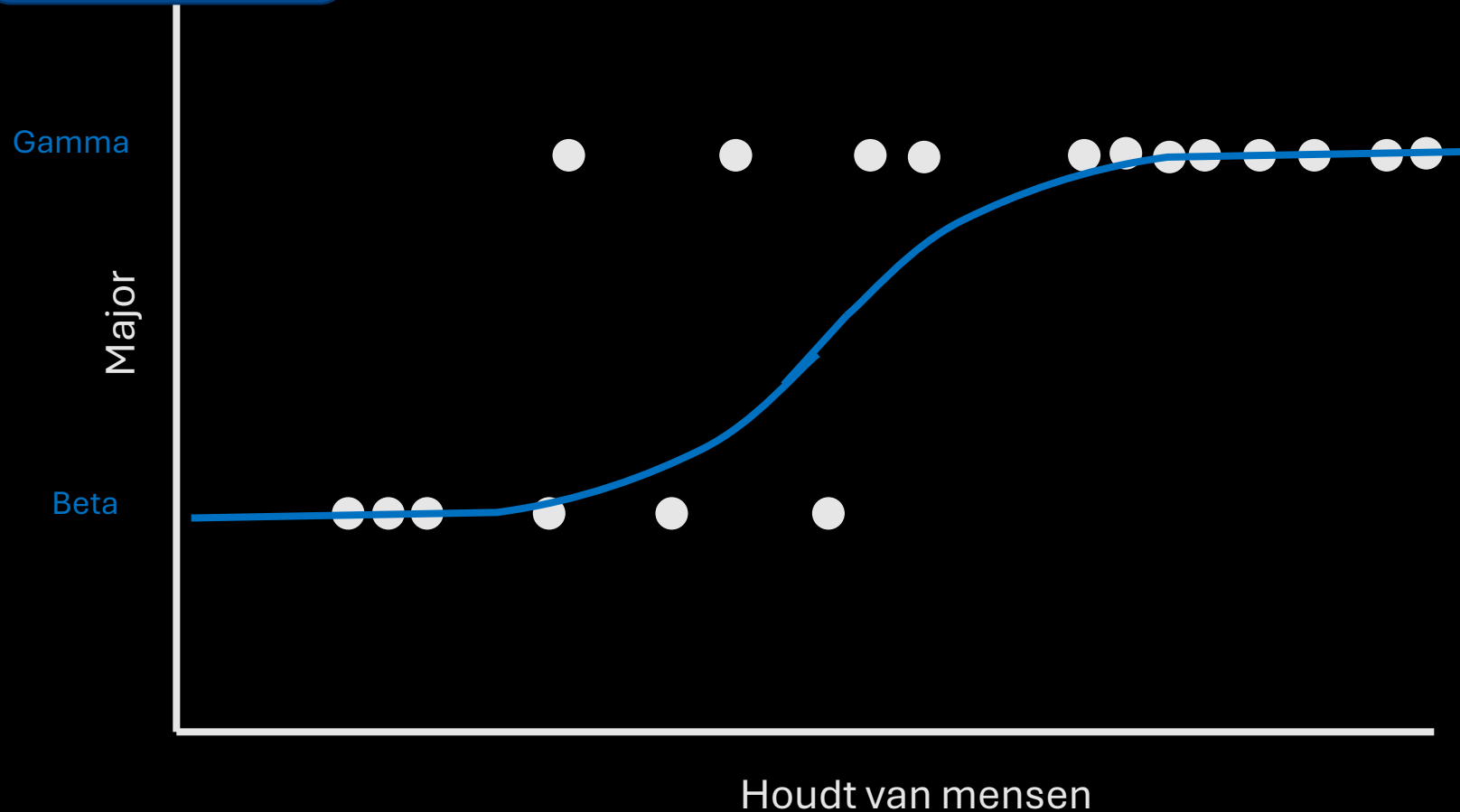
$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

$$Y_i = (\beta_0 + \beta_2 X_{2i}) + (\beta_1 + \beta_3 X_{2i}) X_{1i} + \varepsilon_i$$

Effect van X1 wordt nu zelf
beïnvloed door X2

LOGISTISCH REGRESSIE: DUMMY ALS AFHANKELIJKE VARIABLE

Op de Y-as:
Kansen (p)



INTERPRETEREN : 3 MANIEREN

1. effect op de logged odds:

$$\ln \frac{P_i}{1 - P_i} = \beta_0 + \beta_1 X_i$$

2. effect op de odds:

$$\frac{P_i}{1 - P_i} = e^{\beta_0 + \beta_1 X_i}$$

3. effect op de voorspelde kansen:

$$P_i = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$$

GENESTE DATA

WAT IS GENESTE DATA?

Hierarchisch gestructureerde data

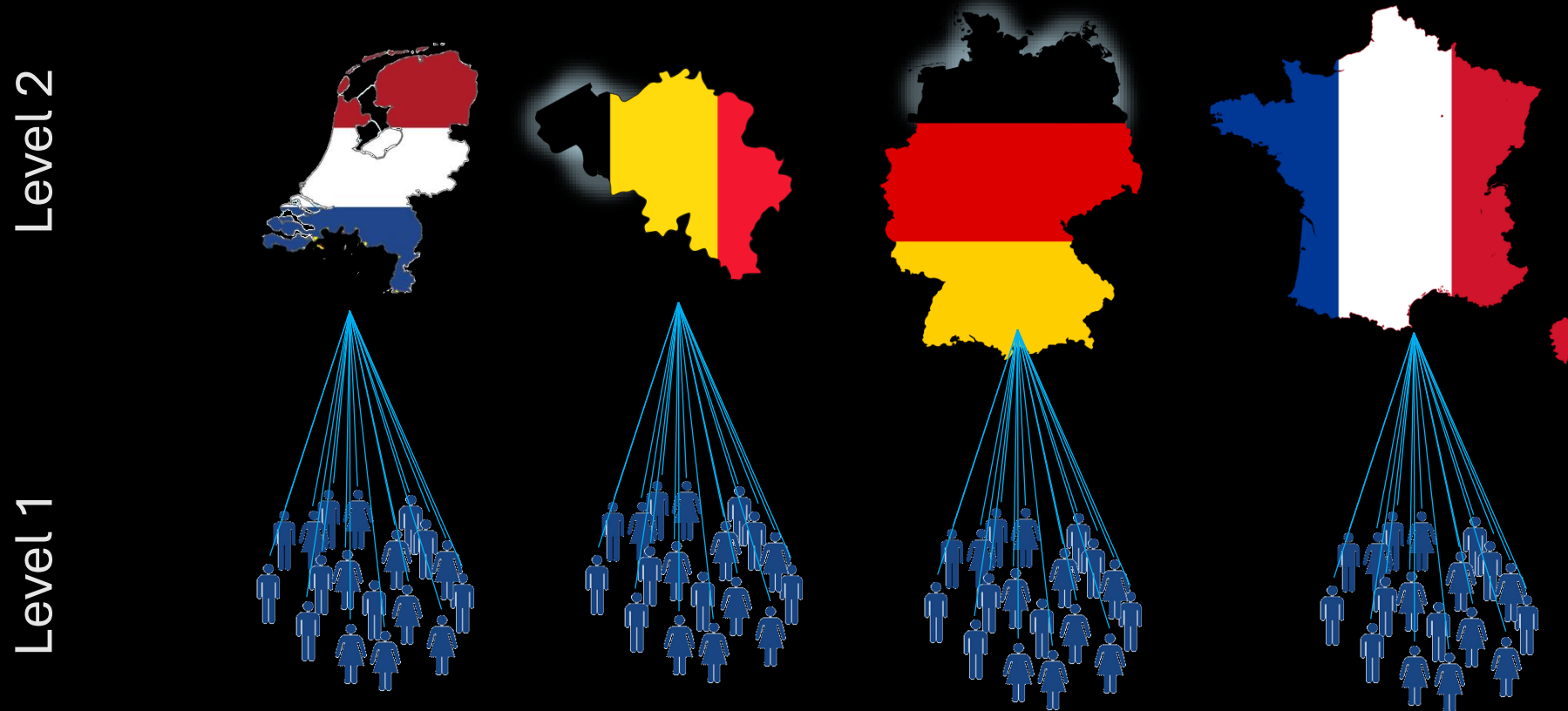
Met andere woorden

observaties zijn gegroepeerd in clusters



VOORBEELDEN VAN GENESTE DATA

- Respondenten binnen landen



VOORBEELDEN VAN GENESTE DATA

- Respondenten binnen landen
- Politici binnen partijen

Level 2



GROE
LINKS

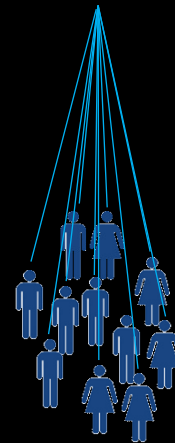
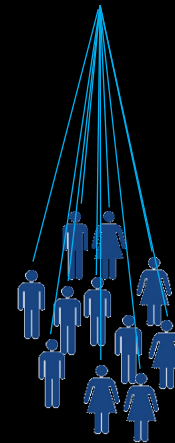
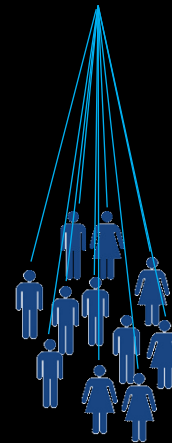
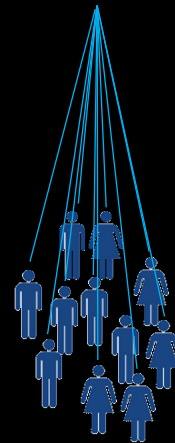
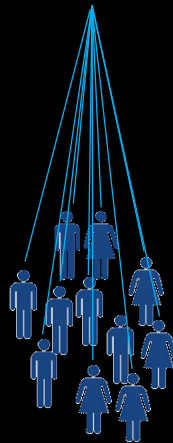
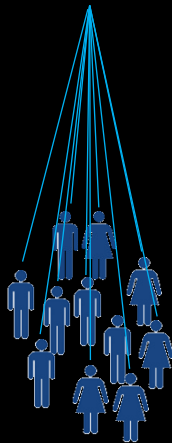
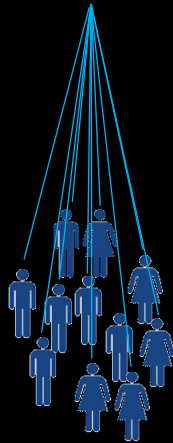


D66

CDA

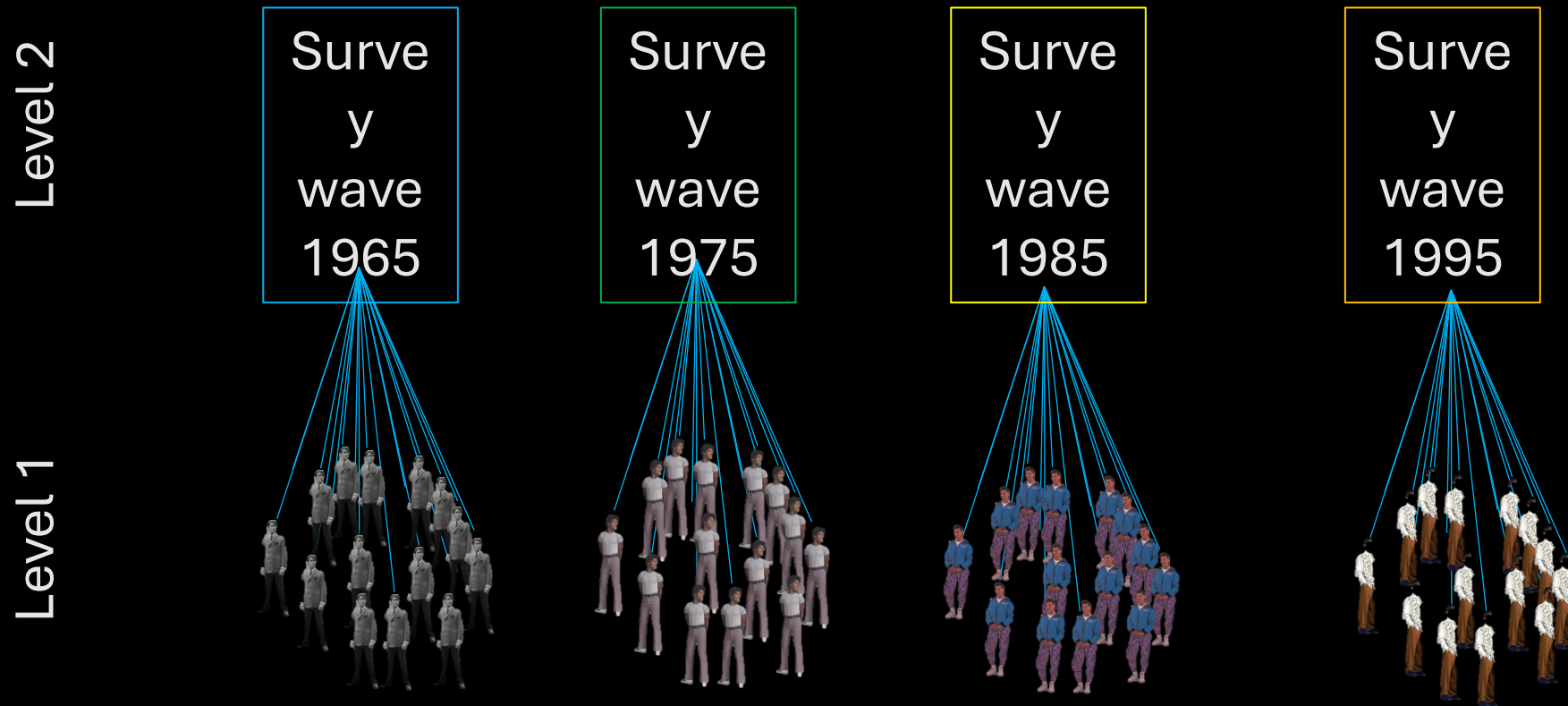


Level 1



VOORBEELDEN VAN GENESTE DATA

- Respondenten binnen landen
- Politici binnen partijen
- Respondenten binnen enquête-rondes



VOORBEELDEN VAN GENESTE DATA

- Respondenten binnen landen
- Politici binnen partijen
- Respondenten binnen enquête-rondes
- Leerlingen in klassen,
klassen in scholen,
scholen in buurten,
buurten in steden,
steden in landen, etc.



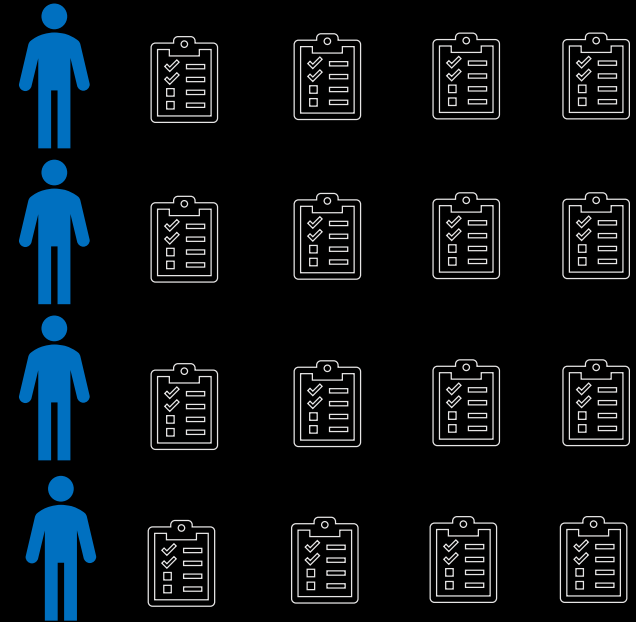
VOORBEELDEN VAN GENESTE DATA

- Respondenten binnen landen
- Politici binnen partijen
- Respondenten binnen enquête-rondes
- Leerlingen in klassen,
klassen in scholen,
scholen in buurten,
buurten in steden,
steden in landen, etc.
- Individuen in huishoudens



VOORBEELDEN VAN GENESTE DATA

- Respondenten binnen landen
- Politici binnen partijen
- Respondenten binnen enquête-rondes
- Leerlingen in klassen,
klassen in scholen,
scholen in buurten,
buurten in steden,
steden in landen, etc.
- Individuen in huishoudens
- Herhaalde metingen binnen respondenten
(panel data, experimentele data etc.)



<https://elmarjansen.nl/os>

OEFENING 1

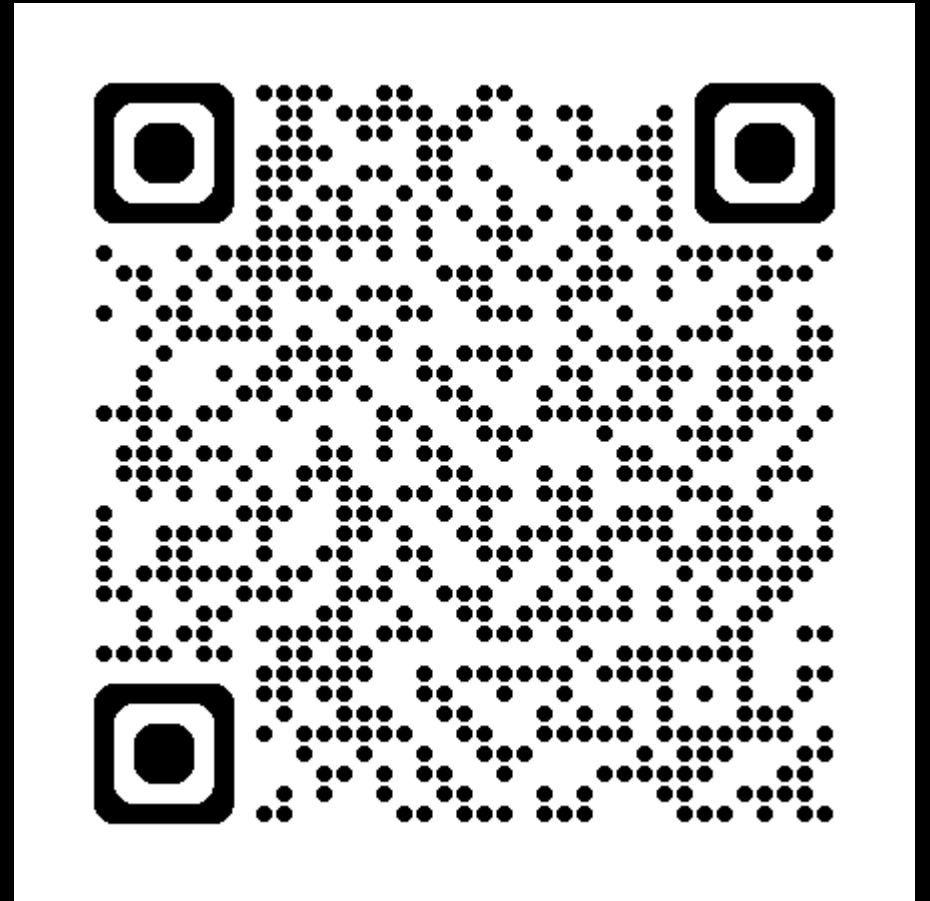
Geneste data:

in 40 straten

steeds 5 mensen ondervraagd

naar gevoel van veiligheid

**Onderzoek: effect van leeftijd op
gevoel van veiligheid op straat**



GENESTE DATA ALS PROBLEEM

Waarom clustering lastig is

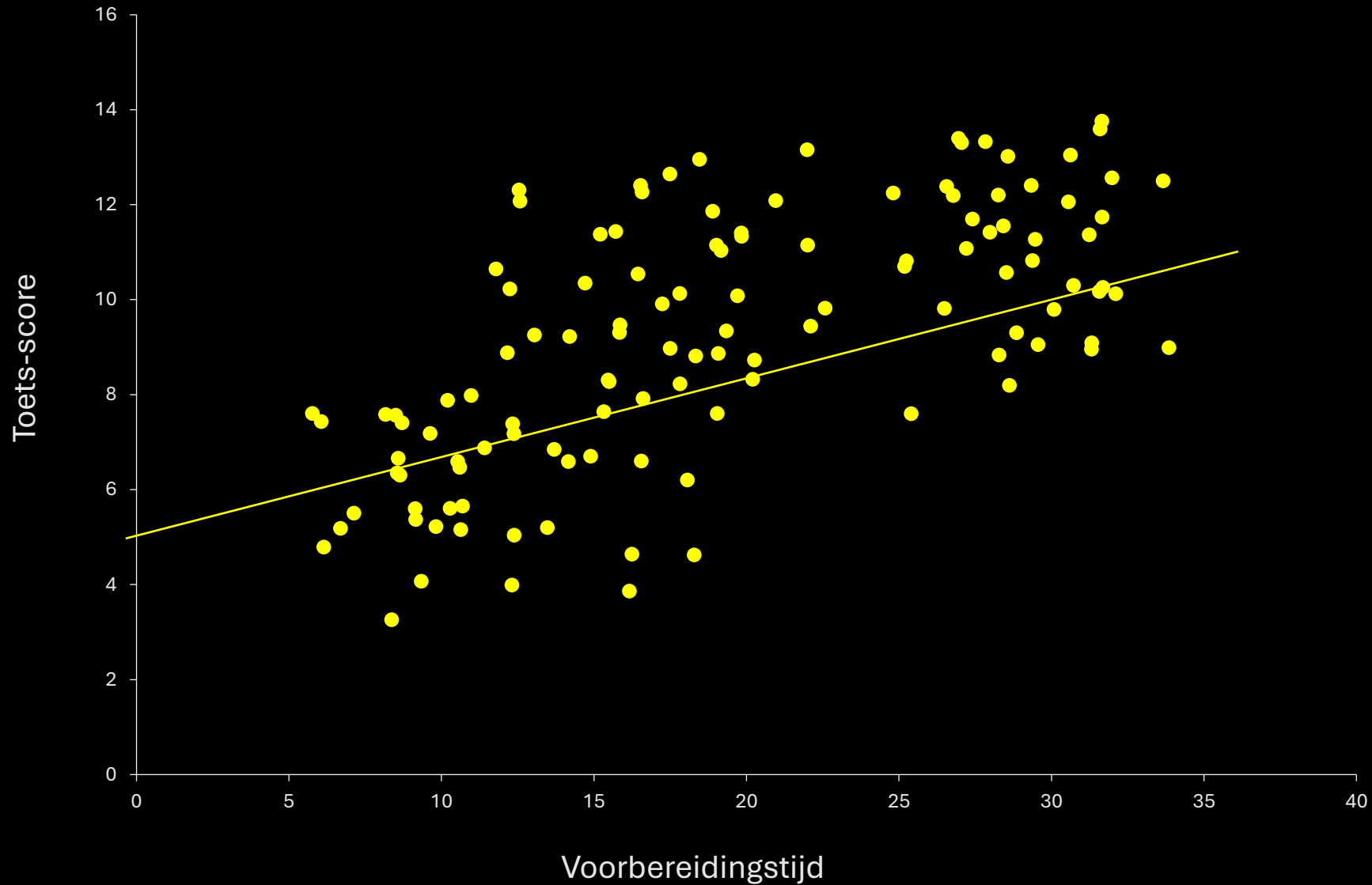
HET PROBLEEM VAN GENESTE DATA...

Afhankelijk van de mate van *correlatie binnen de clusters*,
is dit **valsspelen**:

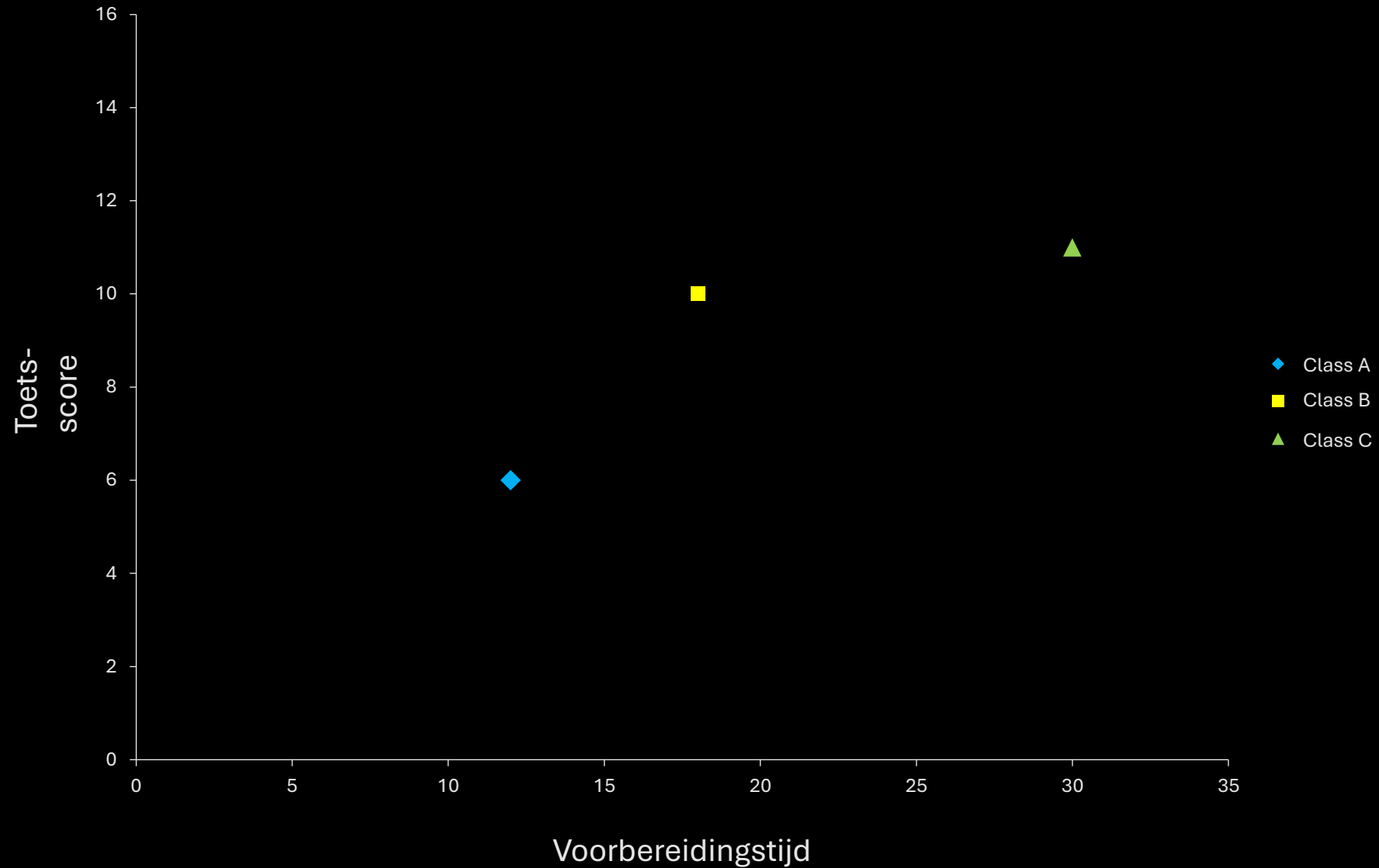
je krijgt **te lage standaardfouten**,
door de N kunstmatig hoog te maken
en dus zijn je gevonden effecten te snel significant

- vooral een probleem als clustering voortkomt uit je onderzoeksontwerp / je manier van dataverzameling
- extra problematisch als clusters niet allemaal even groot zijn

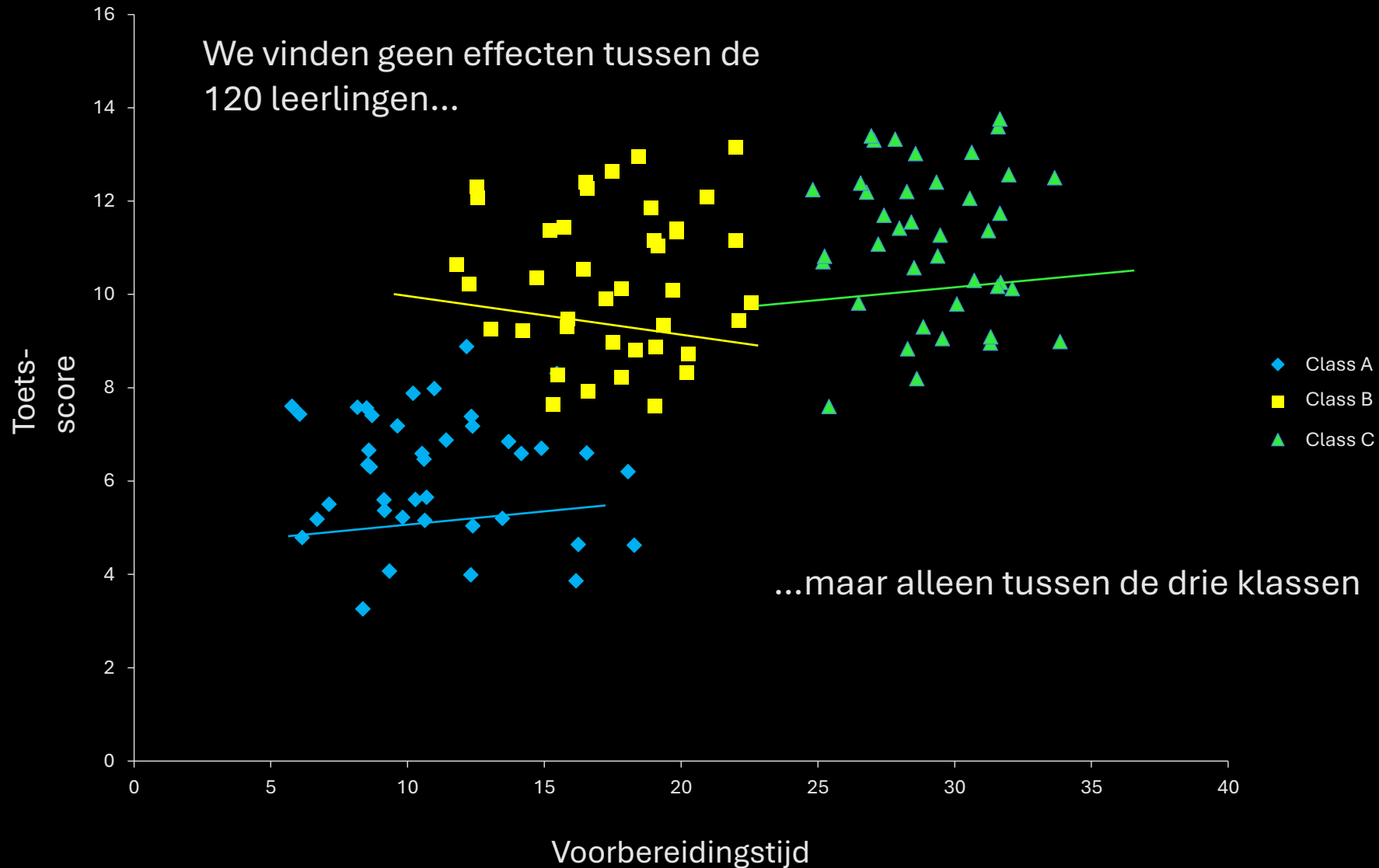
LEERLINGEN UIT DRIE KLASSEN



GEMIDDELDEN VAN DRIE KLASSEN



LEERLINGEN UIT DRIE KLASSEN



DANGER

8 GEVAREN VAN REGRESSIE

DANGER

DANGER!!

1



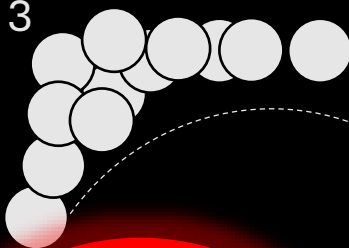
Schijnverband

2



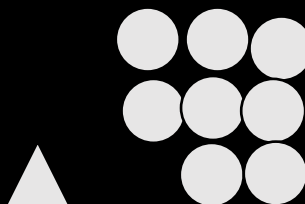
Wederkerigheid /
Simultaniteit

3



Non-Lineairiteit

4



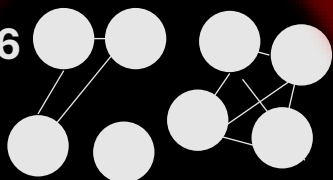
Extreme waarden

5



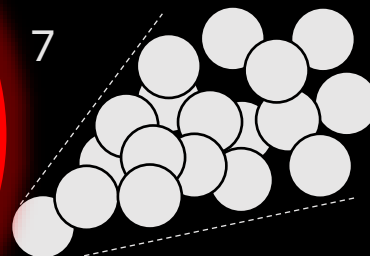
Multicollineariteit

6



Niet onafhankelijke
residuen

7



Heteroskedasticiteit

8

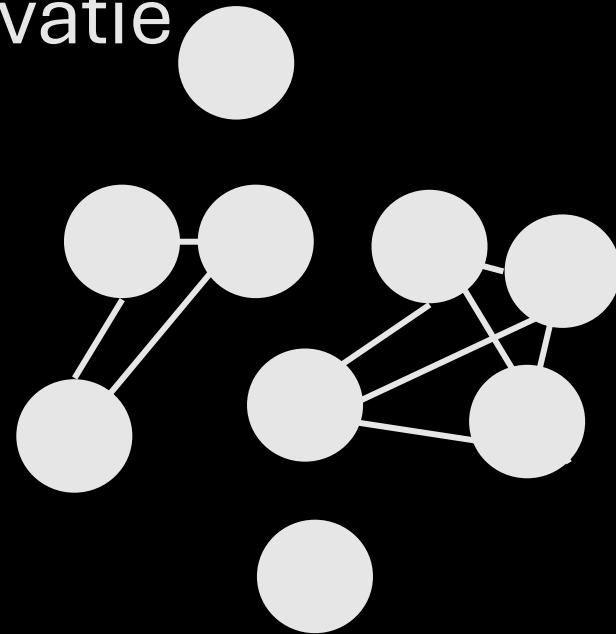


Non-normality
of errors

NIET ONAFHANKELIJKE RESIDUEN

Residu van de ene observatie
mag geen informatie geven
over het residu van een andere observatie

Anders gezegd: een stel observaties mag niet,
door de manier waarop de steekproef is
getrokken, meer op elkaar lijken dan andere
observaties



WAAROM IS DAT EEN PROBLEEM?

Geen willekeurige steekproef meer

als observaties binnen een cluster
meer op elkaar lijken...

brengt een nieuwe observatie
geen volledig nieuwe informatie meer



GEVAAR 6

NIET-ONAFHANKELIJKE RESIDUEN

De gegevens moeten uit een echte aselechte steekproef komen

Alle observaties moeten dus *onafhankelijk* van elkaar zijn

Technische definitie: het residu van de ene observatie mag nooit al iets prijsgeven over het residu van een andere variabele

Gevaar

Onjuiste (te lage) standaard-fouten: onderschatting van onzekerheid

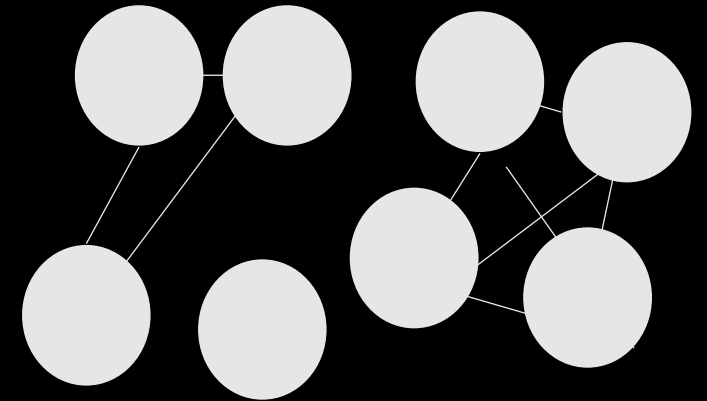
Opsporen van problemen

Nadenken! Geen statistische manier om achter te komen

Oplossingen? Niet met normale (OLS) -regressie

Andere methoden:

- Tijd-series
- Multilevel-analyse
- Paired samples T-test
- Dummies voor groepen ("Fixed effects model")





OPLOSSEN MET DUMMIES

Fixed Effects-model

FIXED EFFECTS MODEL

- **Oplossing:** voeg dummies toe voor alle clusters
 - Je *controleert* zo voor je clusters
 - Met andere woorden: je bestudeert het effect “constant houdend voor” de clusters
 - Variatie *tussen de clusters* wordt op deze manier helemaal niet meer meegenomen

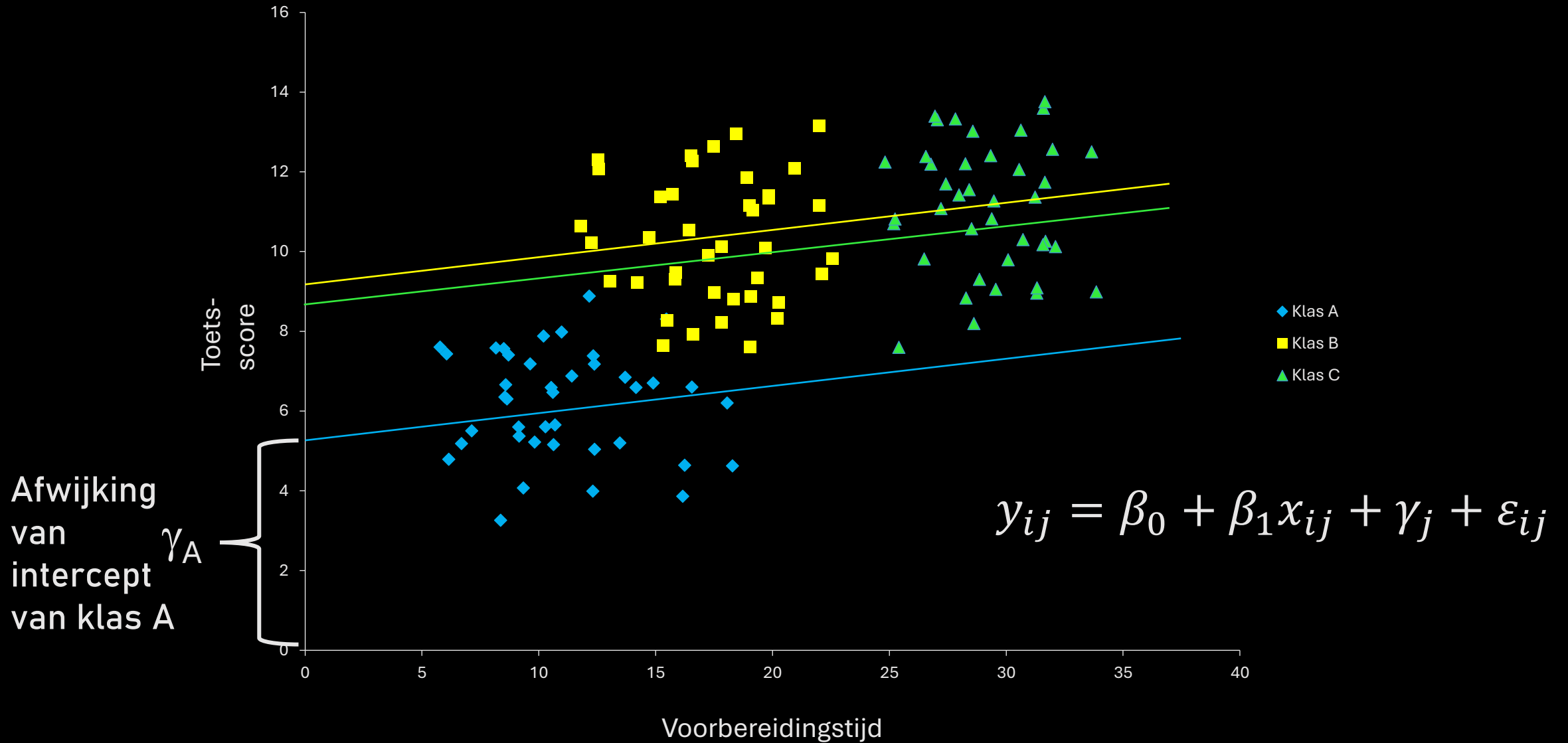
$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 cluster1_j + \beta_3 cluster2_j + \dots + \varepsilon_{ij}$$

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \sum_{k=1}^K \gamma_k cluster_{kj} + \varepsilon_{ij}$$

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_j + \varepsilon_{ij}$$

i is individuele observatie
j is cluster
K is aantal clusters

LEERLINGEN UIT DRIE KLASSEN



<https://elmarjansen.nl/os>

OEFENING 2

Geneste data:

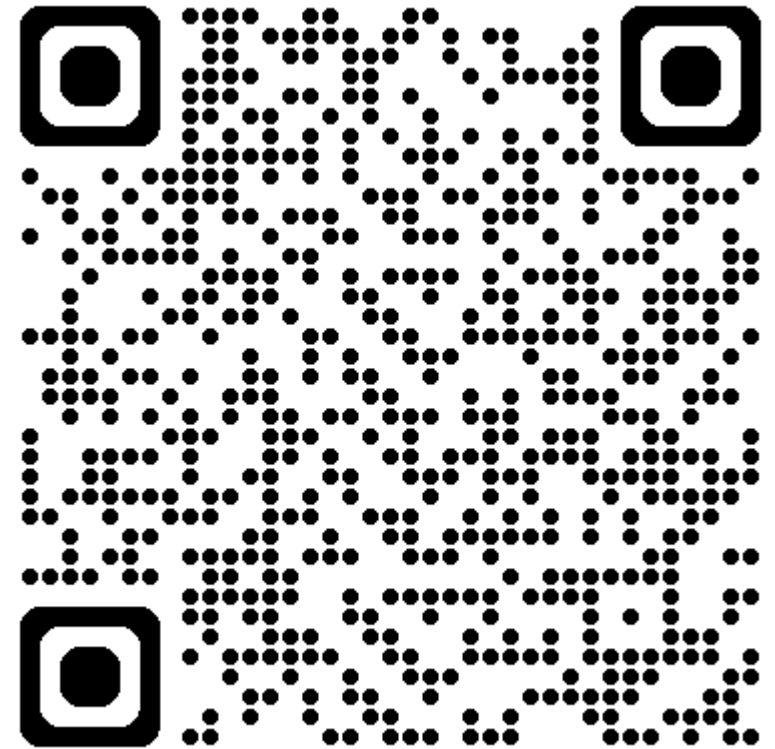
in 40 straten

steeds 5 mensen ondervraagd

naar gevoel van veiligheid

**Onderzoek: effect van leeftijd op
gevoel van veiligheid op straat**

Met dummy-controle voor straat



FIXED EFFECTS ALS BOTTE (MAAR EFFECTIEVE) BIJL



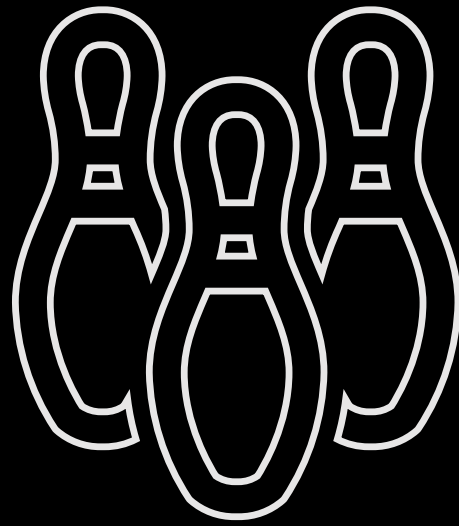
Voordelen

- Heel **veilige** (conservatieve) aanpak: alle variatie tussen clusters weg-gecontroleerd
- Is **eenvoudig** model: kan met “gewone” OLS-regressie door dummies toe te voegen
- Op zijn best bij **klein aantal clusters** of wanneer clusters “bekende” categorieën zijn

Nadelen

- Aanpak **kost veel power**
 - je gebruikt groot deel van de informatie niet: alle variatie *tussen clusters* wordt weggezogen door dummies
 - door al die controlevariabelen neemt de onzekerheid toe
 - standaardfouten worden dus groter
- Weinig effectief als **clusters klein** zijn
- Je kunt **niets meer modelleren** of verklaren op het hogere-niveau

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_j + \varepsilon_{ij}$$



OPLOSSEN MET RANDOM EFFECTS

Mixed-effects-model (oftewel: multilevel-analyse)

RANDOM / MIXED EFFECTS MODEL (“MULTILEVEL”)

- **Oplossing:** voeg een *residu* toe op het niveau van de clusters
 - We zien nu de *clusters* ook als een willekeurige steekproef
 - We laten het model niet alle cluster-afwijkingen schatten met dummies, maar we schatten in het algemeen de *variatie* tussen clusters
 - Deze variatie *tussen clusters* kunnen we eventueel verder verklaren met onafhankelijke variabelen

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + \varepsilon_{ij}$$

i is individuele observatie
j is cluster

LEERLING-VERGELIJKING



Intercept in klas j Effect van oefentijd in klas j

$$score_{ij} = \beta_{0j} + \beta_{1j} tijd_{ij} + \varepsilon_{ij}$$

Leerling i in klas j Oefentijd van leerling i in klas j Residu voor leerling i in klas j

KLAS-VERGELIJKINGEN

$$score_{ij} = \beta_{0j} + \beta_{1j}tijd_{ij} + \varepsilon_{ij}$$



algemene intercept
voor alle klassen

afwijking (residu) van class
j tov. algemene intercept

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Intercept in
klas j

$$\beta_{1j} = \gamma_{10}$$

Effect van
oefentijd in
klas j

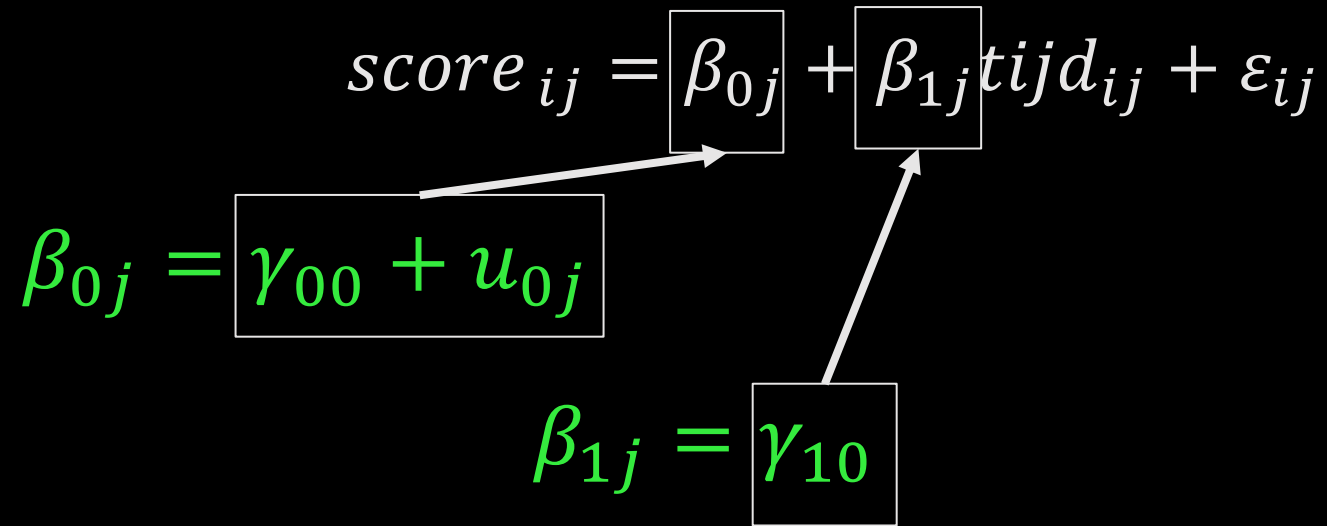
effect is gelijk in
alle klassen

GECOMBINEERDE VERGELIJKING

$$score_{ij} = \boxed{\beta_{0j}} + \boxed{\beta_{1j}} tijd_{ij} + \varepsilon_{ij}$$

$\beta_{0j} = \boxed{\gamma_{00} + u_{0j}}$

$\beta_{1j} = \boxed{\gamma_{10}}$



$$score_{ij} = \gamma_{00} + \gamma_{10} tijd_{ij} + u_{0j} + \varepsilon_{ij}$$

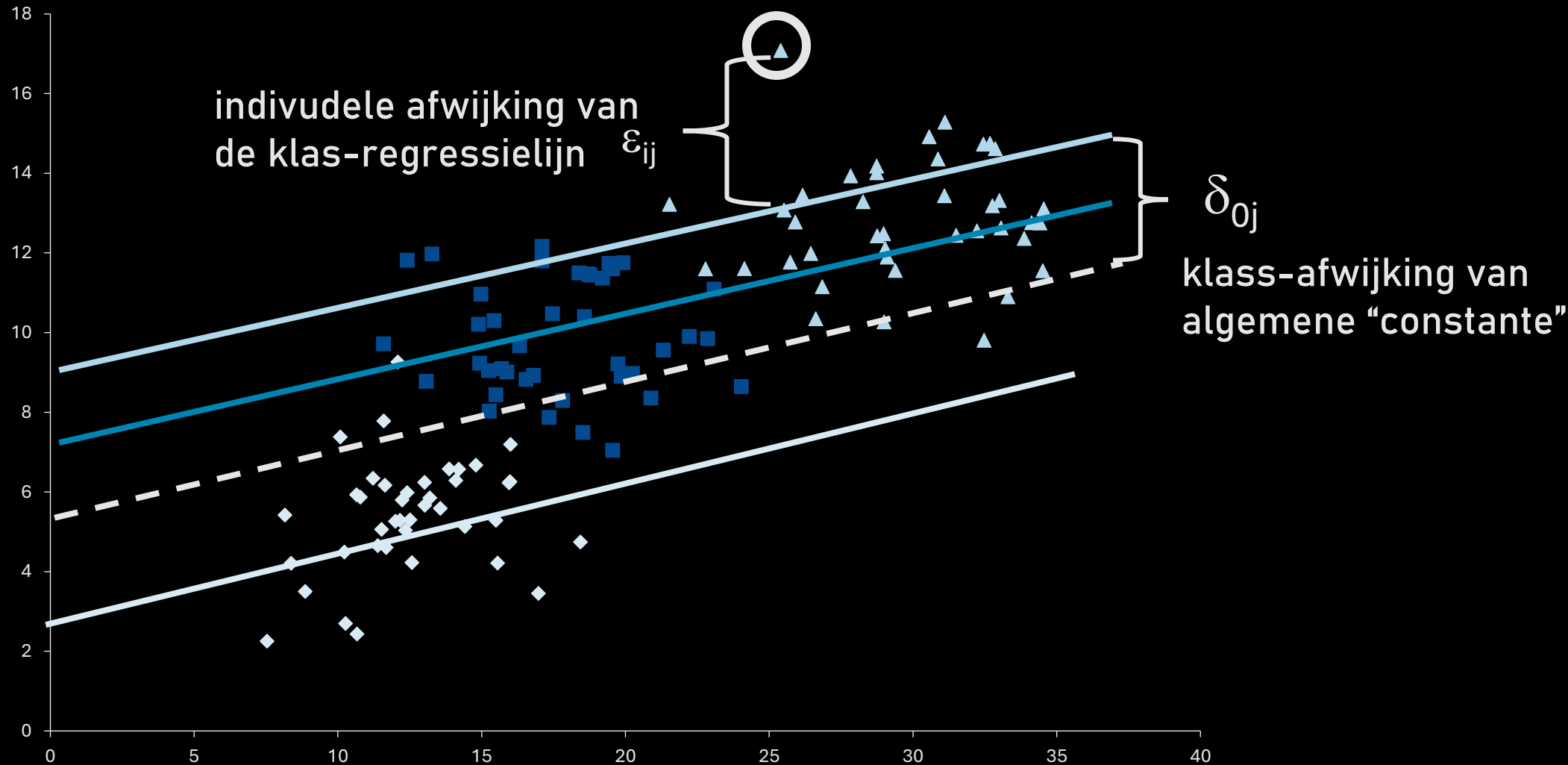
GECOMBINEERDE VERGELIJKING

The diagram illustrates a mixed-effects model equation with four terms, each linked to a descriptive box by a green arrow:

- residu klas j** points to u_{0j}
- residu leerling i in klas j** points to ε_{ij}
- algemene intercept voor alle klassen** points to γ_{00}
- algemeen effect van oefentijd** points to γ_{10}

$$score_{ij} = \gamma_{00} + \gamma_{10}tijd_{ij} + u_{0j} + \varepsilon_{ij}$$

RANDOM INTERCEPT-MODEL



$$score_{ij} = \gamma_{00} + \gamma_{10}t_{ij} + u_{0j} + \varepsilon_{ij}$$

IN R EN SPSS

In R met `lme4` package

- er zijn nog talloze andere packages voor multilevel / mixed effects, maar deze is het eenvoudigst en meest gebruikt

In SPSS met “Mixed Models”

- is even beetje wennen aan het schermpje :)

Geschat worden:

- de *waarden* van γ_{00} en γ_{10}
- de *variantie* (of s.d.) van u_{0j} en ε_{ij}

$$score_{ij} = \gamma_{00} + \gamma_{10}tijd_{ij} + u_{0j} + \varepsilon_{ij}$$

<https://elmarjansen.nl/os>

OEFENING 3

Geneste data:

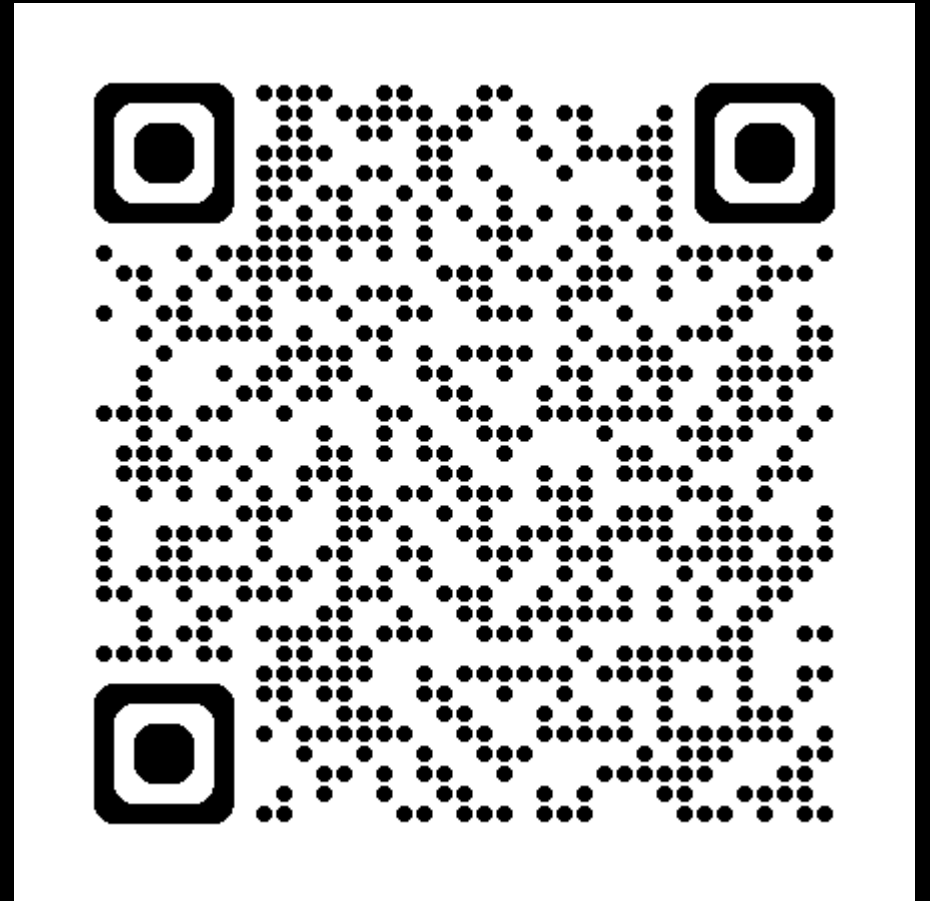
in 40 straten

steeds 5 mensen ondervraagd

naar gevoel van veiligheid

**Onderzoek: effect van leeftijd op
gevoel van veiligheid op straat**

Met random intercept voor straat



MULTILEVEL EFFECTS ALS SUBTIELERE (EN GEVOELIGERE) OPLOSSING



Voordelen

- Kost **weinig power**: er wordt maar één parameter geschat (de variatie van u_{ij})
- Is **veelzijdig**: je kunt nog steeds verklarende variabelen op alle niveaus meenemen
- Op zijn best bij **groot aantal kleine clusters**
- Modelleert ook de **onzekerheid** doordat clusters ook “maar” een steekproef zijn

Nadelen

- Ongeschikt bij **klein aantal clusters** ($N < 10$)
- Aanname: clusters zijn **willekeurige steekproef**
- Vereist **complexere modellen** dan OLS (die vaak ook technische problemen geven bij het schatten)

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + u_{0j} + \varepsilon_{ij}$$

RANDOM SLOPE TOEVOEGEN

Mixed-effects-model (oftewel: multilevel-analyse)

RANDOM / MIXED EFFECTS MODEL (“MULTILEVEL”)

- We kunnen nu ook de *helling* (slope) laten variëren tussen clusters

LEERLING-VERGELIJKING



Intercept in klas j Effect van oefentijd in klas j

$$score_{ij} = \beta_{0j} + \beta_{1j} tijd_{ij} + \varepsilon_{ij}$$

Leerling i in klas j Oefentijd van leerling i in klas j Residu voor leerling i in klas j

KLAS-VERGELIJKINGEN

$$score_{ij} = \beta_{0j} + \beta_{1j}tijd_{ij} + \varepsilon_{ij}$$



algemene intercept
voor alle klassen

afwijking (residu) van klas j
tov. algemene intercept

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Intercept in
klas j

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

Effect van
oefentijd in
klas j

algemeen effect van
oefentijd

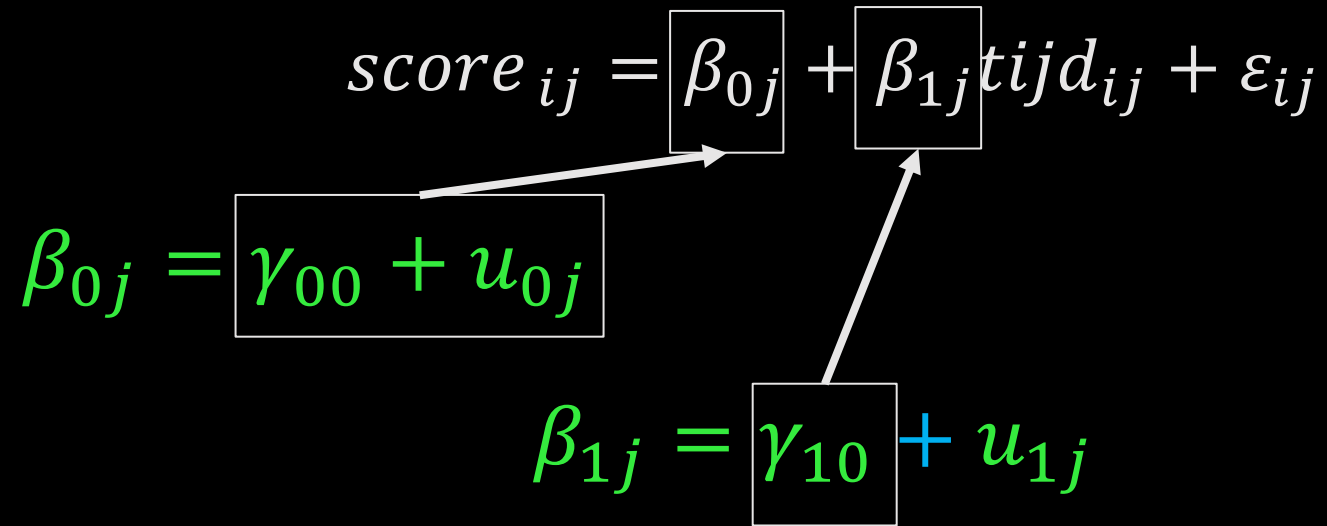
afwijking (residu) van effect van
oefentijd in klas j

GECOMBINEERDE VERGELIJKING

$$score_{ij} = \boxed{\beta_{0j}} + \boxed{\beta_{1j}} tijd_{ij} + \varepsilon_{ij}$$

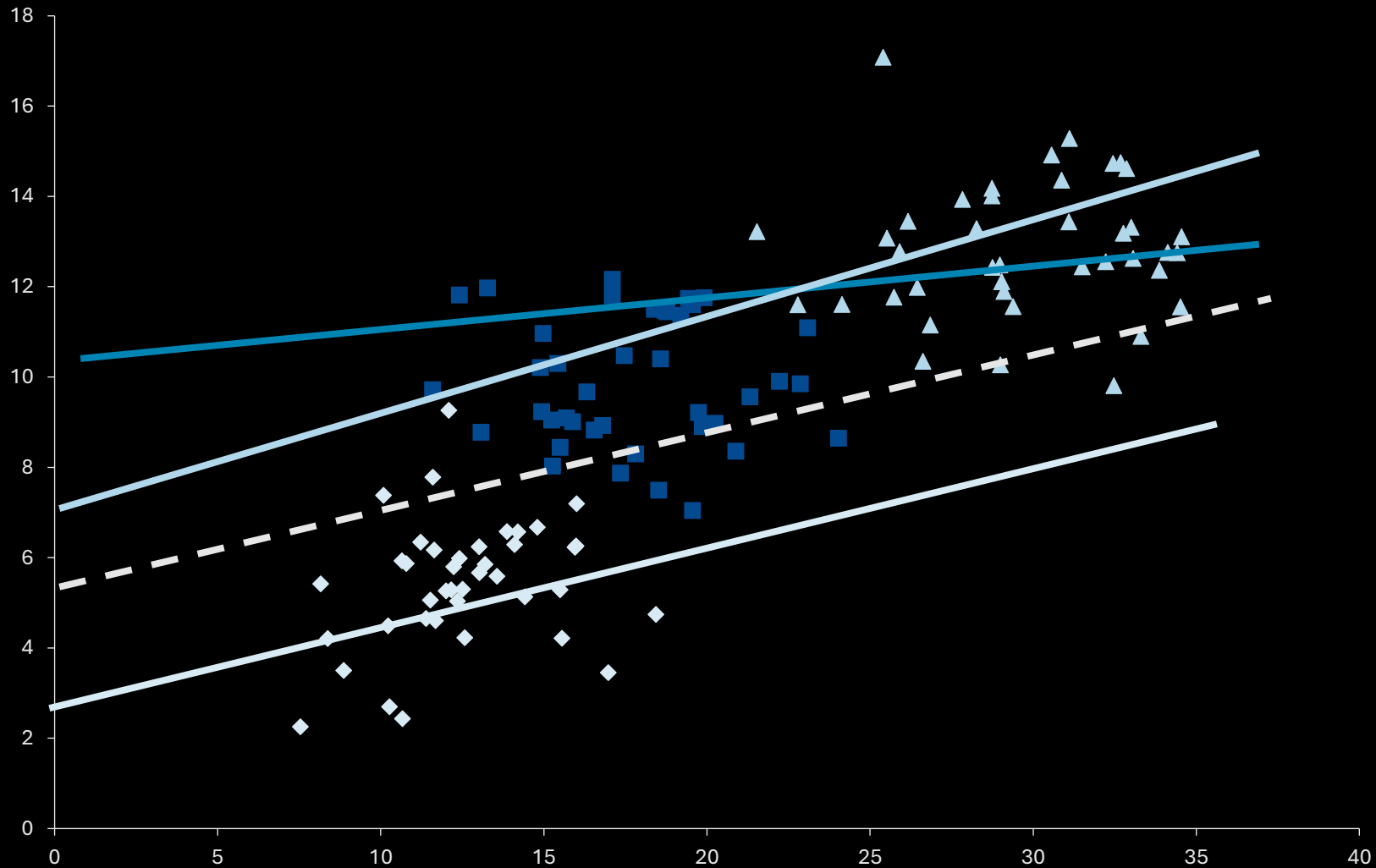
$\beta_{0j} = \boxed{\gamma_{00} + u_{0j}}$

$\beta_{1j} = \boxed{\gamma_{10}} + u_{1j}$



$$score_{ij} = \gamma_{00} + \gamma_{10} tijd_{ij} + u_{0j} + u_{1j} tijd_{ij} + \varepsilon_{ij}$$

RANDOM INTERCEPT-MODEL



$$score_{ij} = \gamma_{00} + \gamma_{10}tijd_{ij} + u_{0j} + u_{1j}tijd_{ij} + \varepsilon_{ij}$$

<https://elmarjansen.nl/os>

OEFENING 4

Geneste data:

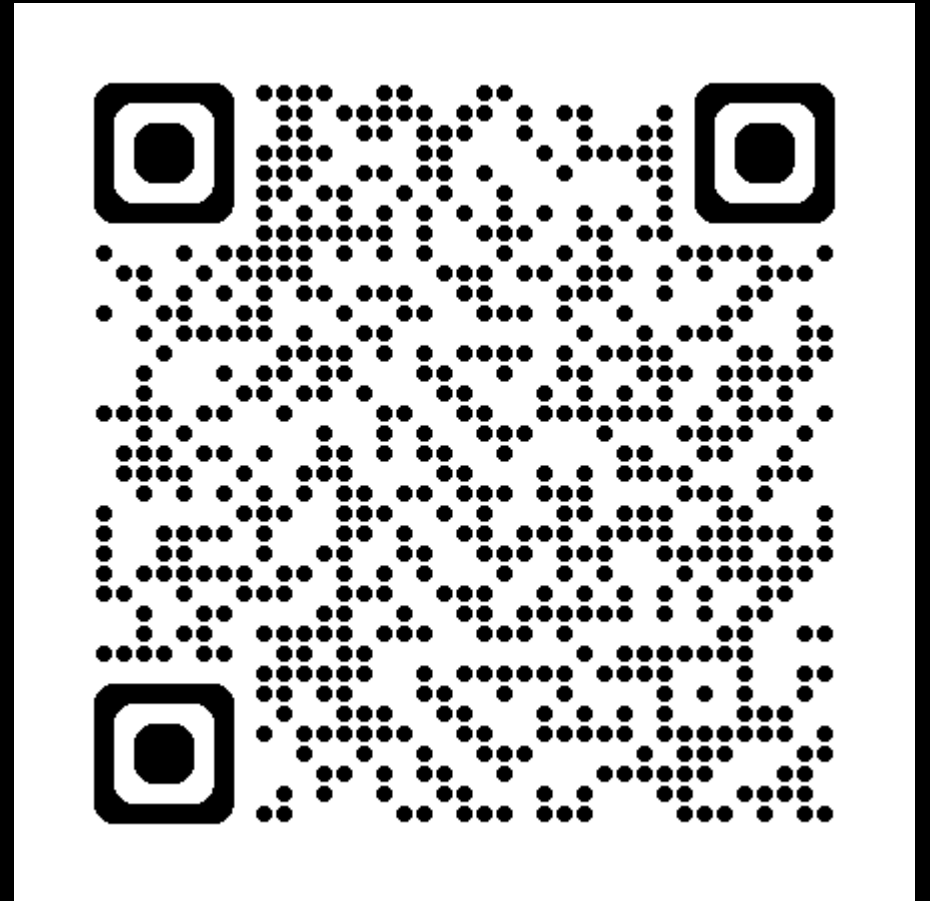
in 40 straten

steeds 5 mensen ondervraagd

naar gevoel van veiligheid

**Onderzoek: effect van leeftijd op gevoel
van veiligheid op straat**

**Met random intercept en slope voor
straat**



DANK JULLIE WEL!!

MULTILEVEL-ANALYSE

18 juni 2024

Training O + S

Elmar Jansen (elmar@elmarjansen.nl)